DRAFT -- To be published in DMT'09 Proceedings
(see http://ngmdb.usgs.gov/Info/dmt/)

A Practitioner's Guide to Managing Geoscience Information

By Jeremy Giles

British Geological Survey
Kingsley Dunham Centre
Keyworth
Nottingham, NG12 5GG
Telephone: +44(0)115 936 3220
Fax: +44(0)115 936 3200
email: jrag@bgs.ac.uk

ABSTRACT

In the United Kingdom, the Natural Environment Research Council manages its scientific data holdings through a series of Environmental Data Centres covering Atmosphere, Bioinformatics, Earth Sciences, Earth Observation, Hydrology, Marine Science, and Polar Science (http://www.nerc.ac.uk/research/sites/data/). Within the Earth Science sector, the National Geoscience Data Centre (NGDC, http://www.bgs.ac.uk/services/ngdc/about.html), a component of the British Geological Survey (BGS), is responsible for managing the geosciences data resource. The purpose of the NGDC is to maintain the national geoscience database and to ensure efficient and effective delivery by providing geoscientists with ready-to-access data and information that is timely, fit for purpose, and in which the user has confidence. The key benefits that NERC derives from this approach are:

- Risk Reduction,
- Increased Productivity, and
- Higher Quality Science.

The paper briefly describes the key benefits of managing geoscientific information effectively and describes how these benefits are realised within the NGDC and BGS.

INTRODUCTION

Geological Survey Organisations (GSO) have three principles resources. These are:

- the expert work force that they employ,
- the facilities to which they have access, and
- the scientific information holdings that they maintain.

This can be likened to a three-legged stool. When all three legs are strong, the stool functions effectively. When any individual leg is weak or missing the whole stool is useless, even if two out of three legs are functioning correctly. Each of these resources needs to be managed. In many GSOs there are professional personnel mangers who manage the staff and there are professional facilities managers to manage the buildings, laboratories, ships, etc. However, the management of the third leg of the stool, information management, is often overlooked. Information is often regarded as 'personal' property and not managed at an organisational asset, so that it is either lost or becomes degraded due to lack of basic maintenance.

The purpose of information management in a GSO is to maintain the national geoscience database and to ensure efficient and effective delivery by providing geoscientists with ready-to-access data and information that is timely, fit for purpose, and in which the user has confidence.

The main drivers for information management are to:

- reduce staff effort in finding data,
- make quality data available to staff and customers,
- facilitate collaboration across and between GSO and other environmental science organisations,
- improve access to the unique information resources,
- inform management decisions, and
- allow Corporate implementation of standards and establish best practice.

The benefits that accrue to a GSO through good information management are:

- Risk Reduction
  - Reduce legislative compliance risk,
  - Reduce litigation risk,
- Increase Productivity, and
- Better Science.

These benefits will be discussed below.

RISK REDUCTION

The information-related risks that a GSO face varies from country to country. The risk level depends upon the legal framework relating to information held by public sector organisations and the risk of litigation that follow the dissemination of information or the provision of advice. The Lofthouse Colliery Disaster is an illustrative example of such a risk. On the 21st March 1973 long-wall advance coal mining was taking place at the Lofthouse Colliery in West-Yorkshire, England. The mining machine cut into unknown old workings of an abandoned colliery called Low Laithes. There was a sudden and catastrophic inrush of water from the old workings that flooded part of the Lofthouse Colliery. Seven miners were killed by the inrush and the colliery was closed shortly afterwards. It transpired at the subsequent enquiry that there was evidence of the existence of the abandoned mine workings held by both the Institute of Geological Sciences (renamed the British Geological Survey in 1984) and other national bodies. However, only a few people in these organisations were aware of the existence of the information and they were unaware of the approaches to their organisations by the mining company that was planning the development of the new face at Lofthouse Colliery. The subsequent report into the disaster by the Chief Inspector of Mines and Quarries recommended the creation of a searchable catalogue of information relating to mining records held by the Institute of Geological Sciences. The Lofthouse Colliery disaster highlighted some of the risks that an organisation is exposed to when it has a limited understanding of the information it holds.

In general the risk can be reduced by:

- Knowing what information is held by a GSO,
- Managing that information as an asset,
- Knowing the quality of the information, and
- Preserving a record of the evidence used to make decisions or to provide advice.

Know the data holdings

It is essential that an organisation know what their data holdings are. It is both good business practice and a key element of a risk reduction strategy. Many countries have introduced some form of Freedom of Information legislation or laws that provide citizens with access to environmental information relating to their communities. Other legislation promotes the re-use and re-purposing of information collected by the public sector organisations, which may include GSOs. Much of this legislation is predicated on the assumption that public sector organisations have a clear understanding of the legacy information that they have collected in the past and the information that they continue to collect at present. It is assumed that an information asset register is available or can be rapidly created to meet the legal requirement. This may not be a trivial task, and may consume considerable resources.

The information asset register requirements of most counties will be met by compiling ISO19115 geospatial metadata (British Standard ISO 19115:2003). However, such an information asset register only provides a top-level view of the data. More detailed metadata will be required to develop a comprehensive understanding of the data holdings. Proactive publication of information asset registers through metadata aggregation services can further reduce risk.

Where resources permit, it is valuable to create digital indexes, or spatially-enabled digital indexes, that show the distribution of individual data points within a dataset. This greatly improves the usability of a data holding and promotes its future re-use and re-purposing.

Managing the Information holdings as an asset

There are two ways to manage information within an organisation: information can be managed as a liability or as an asset. The first way is to manage organisational information as a liability. Nobody wants to take responsibility for the liability, and managers are reluctant to invest time and budget. Information is stored in the cheapest possible ways and is difficult to access. This leads to a downward spiral resulting in a loss of control of the organisations information resource. The alternative approach is to manage information as an organisational asset and to use standard asset management approaches. This involves assigning management responsibilities with clear resources and goals.

Once information assets have been identified, their active management is essential. The phrase "information entropy" was coined by Michener and others (1997) to describe the tendency for stored information to become more disordered over time. The British Geological Survey has adopted an asset-based approach to information management. All datasets have an ISO 19115 metadata record created for the dataset. These metadata records are used as the organisation’s Information Asset Register. The metadata includes the name of the manager who is responsible for the dataset. Their first responsibility is to produce a data management plan for the dataset and to indicate the resources required. Where appropriate, detailed data management procedures are developed to ensure that the dataset is properly maintained and developed. They are also responsible for working with the Intellectual Property Rights manager to ensure that these rights are understood and protected.

Once there is a clear understanding of the information assets of an organisation, decisions can be made about resource allocations. The U.S. Department of Transportation has developed an Asset Management Primer (U.S. Department of Transportation, 1999). This is a process for determining investments and priorities for the management of physical assets for which it is responsible (e.g. bridges). With a little adaption, these processes can be transferred from physical infrastructure assets to information assets (see Figure 1). Typical questions that should be asked include:

- What is the goal of managing the information asset?
- What is the purpose of the information asset?
- What is the quality of the information asset?
- What is the lineage of the information asset?
- How can we preserve the information asset?

- How often is the information asset used?
- What is the cost of preserving the information asset?
- What are the consequences of not maintaining the information asset?
- What is the priority of the individual information asset?

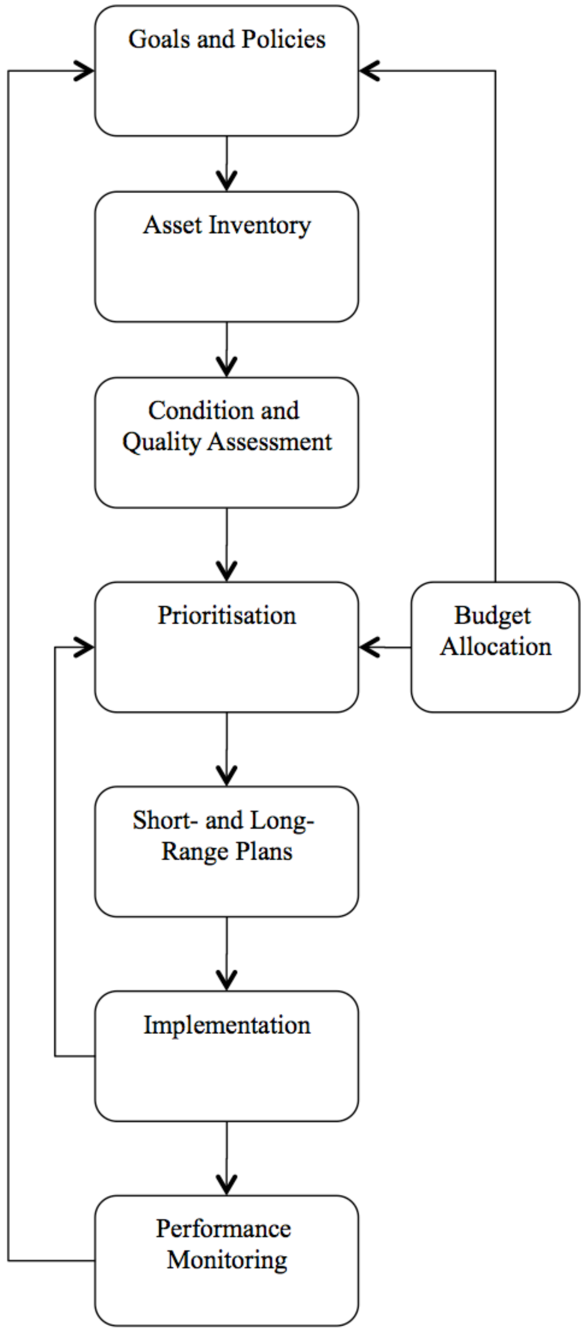

Figure 1. Information Asset Management Primer (from U.S. Department of Transportation, 1999).

There are risks associated with holding information assets that others will re-use and re-purpose. Given that quality is commonly defined as “fit for purpose”, then re-use and re-purposing of information assets will inevitably raise quality issues. Having a clear and well documented, process by which management decisions are made helps to reduce the exposure to such risks.

The discipline of records management is informative for data management practitioners. The practise of data managers has been to retain everything by default. This is rapidly building an unsustainable legacy that will require addressing in the future. A records management practice of retention scheduling and review with these options is very attractive:

i. disposal,
ii. retention for a further period with another review at the end, and
iii. selection for permanent archiving.

Improve quality

There are two elements to improving the quality of an information asset:

- documenting the quality through accurate metadata; and
- addressing known errors.

Metadata is a rich tool. It does so much more than just aiding discovery and identification of datasets. The true purpose of metadata is to allow a potential user of a dataset to assess whether it is fit for their intended purpose. Feineman (1992) identified eight dimensions of data management. These are:

- Accessibility
- Security
- Timeliness
- Accuracy
- Completeness
- Fidelity
- Lineage
- Quality.

These eight dimensions naturally fall into two groups: data management and data quality. The data management dimensions are accessibility, security, and timeliness, whilst the remaining five dimensions relate to data quality. Feineman's ideal for a high quality dataset was one that had exceptional completeness, accuracy, fidelity, and lineage. A comprehensive metadata record allows a potential user to make an accurate assessment of the quality of the datasets. As part of the metadata, a description of the accuracy is important. This should take the form of error limits, and where there are known errors these should be addressed and corrected.

Preserve the evidence

GSOs produce a range of information-based products and services. These include reports, maps, models, geographical information systems, databases, etc. These are used by other organisation to make decisions, develop policy, or make commercial decisions. For this reason it is critical that GSOs preserve the information from which their products and services have been created. It is quite possible for advice provided decades beforehand to be questioned. GSOs therefore need to be able to re-examine the data sources and information used to prepare past products and services. The legal costs of defending past decisions can be considerable. The defence costs can increase considerably if the original data or information cannot be found or if provenance of the evidence is disputed. In terms of risk reduction, it is well worth ensuring that the evidence is preserved in an appropriate records management system.

It is worth noting that the risks associated with digital data appear higher than for paper records. Peritz (1986) noted:

"*…the presumption of trustworthiness (of digital data) simply carries too much weight...*".

Whilst Tarter (1992) noted:
"(the) myth of machine infallibility seems to create a demand for higher standards of quality for machine readable data than for traditionally distributed information."

It appears that once data has become digital it somehow is more trustworthy than mere analogue records. This may be a passing phenomenon, but it should encourage the custodians of digital data to manage its quality carefully.

INCREASED PRODUCTIVITY

Ready access to quality information is essential to scientists. If this is not available then scientists will spend time and effort searching for existing information and improving information quality when they have found it. In the worst case, they may expend resources reacquiring information that already exists. This is not a good use of their time or money. Discovery of information resources can be improved by creating appropriate metadata, a function that can be facilitated by junior staff. Many basic quality checking or quality improving operations can be automated or performed by junior staff, freeing scientist to add value to the information and create knowledge.

Various estimates have been made and studies conducted to quantify the effort that scientists expend in searching for and improving information quality. Two examples are cited below.

Peebler (1996) made the following observation:
"Lack of basic data integration costs the average E&P professional a considerable amount of time. According to various estimates geoscientists and engineers spend from 20% to 30% of their total project time searching for, loading, and formatting data. Obviously, significant productivity gains are still locked up in organizations that do not have level one integration."

In 2002 Shell International undertook a study showing that for New Frontiers Areas, Shell staff spent their time as follows:

- Finding data - 53%
- Archiving data - 9%
- Documenting the data - 15%
- Interpreting (adding value) - 23%.

Shell set goals to raise the time spent interpreting the data (adding value) to 46%, by reducing the time to find data to 30% (Source: BGS Strategy 2009).

Both studies suggest there is considerable potential to increase a scientist's productivity if they have ready access to quality information.

BETTER SCIENCE

Good information management contributes to improving the quality and reliability of scientific outputs in a number of ways:

- Preserving the evidence
- Re-use of existing information
- Re-purposing of existing information.

Information collected during a scientific research project forms a key component of the record of that project. Should the results of the study be questioned in the future, the preserved record of the project can be reviewed to ensure that the conclusions and recommendations of the project remain valid or whether a reinterpretation is justified.

The information collected during a project can be reused at a later date. For example information could be collected for a study area during a research project and preserved after the project is concluded. At a later date a new project undertakes a regional study, which reuses the data collected in a previous project and undertakes additional collection of information in other parts of the region. The opportunity to reuse existing information reduces the cost of the subsequent study.

Information collected for a specific purpose by a research project can be used for an entirely different purpose that was not envisaged when the original information was acquired. For example, the British Geological Survey routinely acquires borehole logs from site investigations; over the course of time well over a million logs have been collected from across Great Britain. The original purpose of these boreholes was to gauge the foundation conditions for a proposed building project. However, information contained within these borehole logs was aggregated to produce a national superficial sediment thickness model.

CONCLUSION

There are a range of significant benefits that can be realised through a well organised and well resourced information management programme. These benefits can only be realised through careful planning and implementation. Key elements of such a programme include:

- Creating metadata that enables the information resource to be discovered and the quality of that resource to be assessed.
- Digital indexes, that may be spatially enabled, created for key resources so that individual items within information can be located.
- Asset-based information practises are adopted so there is a clear plan for investment in specific information resources both in terms of business need and long-term preservation.

REFERENCES

British Geological Survey, 2009, British Geological Survey Strategy 2009-2014 -- Applied geosciences for our changing Earth: Nottingham, British Geological Survey document.

BS ISO 19115:2003, Geographic Information – Metadata: London, British Standards Institution, 2003.

Feineman, D.R. 1992, Data Management: Yesterday, Today and Tomorrow: Presentation to the PETEX ’92 Conference, 25 November 1992, London.

Michener, W.K., Brunt, J.W., Helly, J., Kirchner, T.B., and Stafford, S.G., 1997, Non-geospatial metadata for the ecological sciences: Ecological Applications, v. 7, p. 330-342.

Peebler, R., 1996, Extended Integration -- The Key To Future Productivity Leap: Oil and Gas Journal, v. 94, no. 21.

Peritz, R., 1986, Computer Data and Reliability: Northwest University Law Review, v. 80, p. 960.

Tarter, B., 1992, Information Liability: New Interpretations for the Electronic Age: Computer/Law Journal, v. XI, p. 484.

U.S. Department of Transportation, 1999, Asset Management Primer: Washington, D.C., U.S. Department of Transportation, Federal Highways Administration, Office of Asset Management.